

Scotland's Rural College

## **Imputation of genotypes from low density (50,000 markers) to high density (700,000 markers) of cows from research herds in Europe, North America, and Australasia using 2 reference populations**

Pryce, JE; Johnston, J; Hayes, BJ; Sahana, G; Weigel, KA; McParland, S; Spurlock, D; Krattenmacher, N; Spelman, RJ; Wall, E; Calus, MPL

*Published in:*  
Journal of Dairy Science

*DOI:*  
[10.3168/jds.2013-7368](https://doi.org/10.3168/jds.2013-7368)

First published: 01/01/2014

*Document Version*  
Publisher's PDF, also known as Version of record

[Link to publication](#)

### *Citation for pulished version (APA):*

Pryce, JE., Johnston, J., Hayes, BJ., Sahana, G., Weigel, KA., McParland, S., Spurlock, D., Krattenmacher, N., Spelman, RJ., Wall, E., & Calus, MPL. (2014). Imputation of genotypes from low density (50,000 markers) to high density (700,000 markers) of cows from research herds in Europe, North America, and Australasia using 2 reference populations. *Journal of Dairy Science*, 97(3), 1799 - 1811. <https://doi.org/10.3168/jds.2013-7368>

### **General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

### **Take down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.



## Imputation of genotypes from low density (50,000 markers) to high density (700,000 markers) of cows from research herds in Europe, North America, and Australasia using 2 reference populations

J. E. Pryce,\*††<sup>1</sup> J. Johnston,§ B. J. Hayes,\*†‡ G. Sahana,# K. A. Weigel,|| S. McParland,¶ D. Spurlock,\*\* N. Krattenmacher,†† R. J. Spelman,‡‡ E. Wall,§§ and M. P. L. Calus##

\*Department of Environment and Primary Industries, Agribio, 5 Ring Road, La Trobe University, Bundoora, VIC 3083, Australia

†Dairy Futures Cooperative Research Centre, 5 Ring Road, La Trobe University, Bundoora, VIC 3083, Australia

‡La Trobe University, 5 Ring Road, La Trobe University, Bundoora, VIC 3083, Australia

§Canadian Dairy Network, Guelph, Ontario, N1K 1E5, Canada

#Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, 8830 Tjele, Denmark

||Department of Dairy Science, University of Wisconsin, Madison 53706

¶Animal & Grassland Research and Innovation Centre, Teagasc, Moorepark, Co. Cork, Ireland

\*\*Department of Animal Science, Iowa State University, Ames 50011

††Institute of Animal Breeding and Husbandry, Christian-Albrechts-University, 24118 Kiel, Germany

‡‡LIC, Private Bag 3016, Hamilton 3240, New Zealand

§§Animal and Veterinary Sciences, Scotland's Rural College (SRUC), Kings Buildings, West Mains Road, Edinburgh, EH9 3JG, United Kingdom

##Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, 8200 AB Lelystad, the Netherlands

### ABSTRACT

Combining data from research herds may be advantageous, especially for difficult or expensive-to-measure traits (such as dry matter intake). Cows in research herds are often genotyped using low-density single nucleotide polymorphism (SNP) panels. However, the precision of quantitative trait loci detection in genome-wide association studies and the accuracy of genomic selection may increase when the low-density genotypes are imputed to higher density. Genotype data were available from 10 research herds: 5 from Europe [Denmark, Germany, Ireland, the Netherlands, and the United Kingdom (UK)], 2 from Australasia (Australia and New Zealand), and 3 from North America (Canada and the United States). Heifers from the Australian and New Zealand research herds were already genotyped at high density (approximately 700,000 SNP). The remaining genotypes were imputed from around 50,000 SNP to 700,000 using 2 reference populations. Although it was not possible to use a combined reference population, which would probably result in the highest accuracies of imputation, differences arising from using 2 high-density reference populations on imputing 50,000-marker genotypes of 583 animals (from the UK) were quantified. The European genotypes ( $n = 4,097$ ) were imputed as 1 data set, using a reference population of 3,150 that included genotypes from 835 Australian and 1,053 New Zealand females, with the re-

mainder being males. Imputation was undertaken using population-wide linkage disequilibrium with no family information exploited. The UK animals were also included in the North American data set ( $n = 1,579$ ) that was imputed to high density using a reference population of 2,018 bulls. After editing, 591,213 genotypes on 5,999 animals from 10 research herds remained. The correlation between imputed allele frequencies of the 2 imputed data sets was high ( $>0.98$ ) and even stronger ( $>0.99$ ) for the UK animals that were part of each imputation data set. For the UK genotypes, 2.2% were imputed differently in the 2 high-density reference data sets used. Only 0.025% of these were homozygous switches. The number of discordant SNP was lower for animals that had sires that were genotyped. Discordant imputed SNP genotypes were most common when a large difference existed in allele frequency between the 2 imputed genotype data sets. For SNP that had  $\geq 20\%$  discordant genotypes, the difference between imputed data sets of allele frequencies of the UK (imputed) genotypes was 0.07, whereas the difference in allele frequencies of the (reference) high-density genotypes was 0.30. In fact, regions existed across the genome where the frequency of discordant SNP was higher. For example, on chromosome 10 (centered on 520,948 bp), 52 SNP (out of a total of 103 SNP) had  $\geq 20\%$  discordant SNP. Four hundred and eight SNP had more than 20% discordant genotypes and were removed from the final set of imputed genotypes. We concluded that both discordance of imputed SNP genotypes and differences in allele frequencies, after imputation using different reference data sets, may be used to identify and remove poorly imputed SNP.

**Key words:** imputation, high-density genotyping

Received August 14, 2013.

Accepted December 3, 2013.

<sup>1</sup>Corresponding author: [jennie.pryce@depi.vic.gov.au](mailto:jennie.pryce@depi.vic.gov.au)

## INTRODUCTION

Recently, interest has grown in combining data from international research populations for genetic analysis, especially for difficult-to-measure traits such as DMI (Banos et al., 2012; de Haas et al., 2012; Veerkamp et al., 2012). Reasons why genotypes and phenotypes from different research organizations are combined include adding statistical power to genome-wide association studies (**GWAS**) and trying to improve the accuracy of genomic prediction. For example, significant putative QTL were detected by Berry et al. (2012) and Veerkamp et al. (2012) by combining data from 4 countries that may not have been observed using data from the individual countries alone.

Genotyping at high density could provide markers that are closer to and therefore in stronger linkage disequilibrium (**LD**) with QTL. This can be used to increase the precision of QTL detection. For example using high-density (**HD**) genotypes [approximately 700,000 (**700K**) SNP], VanRaden et al. (2013) refined the precision of a QTL that had a large effect on several traits and is positioned on chromosome 18, which was previously reported with less accurate precision using lower density [approximately 50,000 (**50K**) SNP] by Cole et al. (2009).

Increasing the size of the reference population by combining data across countries, using high-density genotypes could be particularly advantageous, because it should facilitate greater accuracy of prediction across less-related populations (de Roos et al. 2009). Harris et al. (2011) found no increase in reliability of genomic breeding values estimated using high-density compared with low-density genotypes; other studies have documented a small increase (Erbe et al., 2012; Su et al., 2012; VanRaden et al., 2013). A larger increase in reliability was observed when Bayesian methods were used for genomic prediction instead of the commonly used genomic BLUP (Erbe et al., 2012; Gao et al., 2013).

A constraint in many research herds is the cost of genotyping, with most currently opting for genotyping SNP at 50K. So, value may exist in predicting the additional SNP on the high-density chip using imputation methods. Imputation is the process of predicting genotypes by finding haplotype segments that are shared between individuals (for a review, see Marchini and Howie, 2010). Some imputation methods also use pedigree information in addition to the LD among markers (e.g., Sargolzaei et al., 2011; VanRaden et al., 2011; Hickey et al., 2012). Imputation can be affected by the genetic composition of the reference population, with fewer imputation errors occurring when close relatives are genotyped at high density (Zhang and Druet, 2010; Mulder et al., 2012). To our knowledge, no published

studies have investigated the effect of using different reference populations genotyped at 700K SNP on the accuracy of imputation.

The overall aim of this study was to derive/provide guidelines on how to prepare and impute genotypes that are combined from multiple sources. Specifically, we investigated (1) whether the imputation of genotypes from comparatively low density (50K SNP) to HD (around 700K SNP) is affected by using different reference populations genotyped at HD, (2) if this approach can be used to identify poorly imputed regions by investigating discordant imputed genotypes of the same data set using 2 different reference populations, and (3) estimate genetic diversity in the 10 research populations included in this study.

## MATERIALS AND METHODS

Ten research herds from 9 countries contributed genotype data from Holstein dairy cattle to this study. The animals were selected for this study because they had phenotypes for DMI.

Data set **EuroAus** consisted of research herds from European and Australasian countries [Australia, Denmark, Germany, Ireland, the Netherlands, New Zealand, and the United Kingdom (**UK**)]. Data set **NAmer** comprised the same research herd from the UK [belonging to Scotland's Rural College (SRUC), Edinburgh, UK] and North American countries: 2 research herds from the United States [belonging to Iowa State University (Ames) and University of Wisconsin (Madison)] and 1 research herd from Canada (University of Alberta, Edmonton, AB, Canada). The UK genotypes were included in both data set EuroAus and NAmer. Having UK animals included in both data sets was useful from a quality control perspective, as imputation results from the 2 data sets could be compared using these animals. Although it would have been interesting to compare imputation using each reference data set to a combined reference data set (of EuroAus and NAmer), this was not possible because of sensitivities around genotype ownership. However, the UK animals could be included in both data sets because the UK is part of the InterContinental Genomic Consortium that includes Canada, the United States, the UK, and Italy. This consortium had a large HD reference population available for imputing genotypes of member countries from 50K to HD.

Genotyping quality control of genotypes from the BovineSNP50 BeadChip (Illumina Inc., San Diego, CA; Matukumalli et al., 2009) abbreviated to 50K and imputation of genotypes from 50K to HD, using the BovineHD Genotyping BeadChip (Illumina Inc.), were undertaken independently in 2 separate data sets



described in Table 1. All the SNP (50K and HD) were mapped to the build UMD 3.1 (Center for Bioinformatics and Computational Biology, University of Maryland, College Park; 2010; [http://cbcb.umd.edu/research/bos\\_taurus\\_assembly.shtml](http://cbcb.umd.edu/research/bos_taurus_assembly.shtml)). In data set EuroAus, 3,884 animals (males and females) were genotyped on the first version of the Illumina 50K SNP chip and 836 using the second version of the Illumina 50L SNP chip; the number of genotypes from females are presented in Table 1. Data editing of 50K genotypes was performed in the batch that was obtained from the laboratory or genotype provider for data set EuroAus and by country for data set NAmr. In data set NAmr, 1,050 animals were genotyped with the first version of the 50K SNP and 685 with the second version (Table 1). Only SNP on version 2 of the BovineSNP50 BeadChip were considered.

The same genotype editing rules were applied within batch (i.e., to the data from each research herd) and then within data sets EuroAus and NAmr. The editing rules that were applied within batch were as follows:

- Overall call rate (across animals) of individual SNP  $\geq 95\%$ ;
- Overall call rate (across SNP) for each animal  $\geq 95\%$ ;
- Illumina GenTrain scores (Gen-Call) of individual SNP were accepted if they were  $>60\%$  ([http://res.illumina.com/documents/products/technotes/technote\\_gencall\\_data\\_analysis\\_software.pdf](http://res.illumina.com/documents/products/technotes/technote_gencall_data_analysis_software.pdf));
- Exclude mitochondrial SNP, Y SNP, unmapped SNP, and mismatched SNP, and so on;
- Exclude SNP with Mendelian inconsistencies.

No edits were made on minor allele frequency (MAF), as it was possible that SNP with low MAF in one batch may be higher in another batch. Within data set NAmr and across batches within data set EuroAus, within-data-set SNP were excluded at positions where 2 homozygotes but no heterozygote were observed, or SNP for which only heterozygotes were observed (across batches). Single nucleotide polymorphisms on the X chromosome were retained as long as they passed quality control.

The same quality control steps and editing rules were applied to both the 50K genotypes and the EuroAus HD set, and then these filtered HD SNP were selected for the NAmr reference population. The HD reference sets comprised 3,150 and 2,018 individuals for EuroAus and NAmr, respectively.

Several programs are available for genotype imputation and most of them impute a high proportion of genotypes correctly (Ma et al., 2013). The software BEAGLE (Browning and Browning, 2009) was chosen

**Table 1.** A summary of genotypes received, the SNP chip used, and the imputation data set that the animals were assigned to

Country of origin	Research herd	Number of animal genotypes received	SNP chip used <sup>1</sup>	50K to HD imputation data set <sup>2</sup>
Australia	Department of Environment and Primary Industries (DEPI; Melbourne, VIC, Australia)	835	HD	N/A
Canada	University of Alberta (Edmonton, AB, Canada)	201	50K v1	NAmr
Denmark	Aarhus University (Tjele, Denmark)	280	50K v2	EuroAus
Germany	Christian-Albrechts-University (Kiel, Germany)	835	50K v1 and v2	EuroAus
Ireland	Teagasc (Moorepark, Co. Cork, Ireland)	549	50K v1	EuroAus
The Netherlands	Wageningen UR Livestock Research (Lelystad, the Netherlands)	1,147	50K v1 (700) and v2 (447)	EuroAus
New Zealand	DairyNZ/LIC (Hamilton, New Zealand)	1,053	HD	N/A
United Kingdom	Scotland's Rural College (SRUC; Edinburgh, UK)	685	50K v2	EuroAus and NAmr
United States (Iowa)	Iowa State University (Ames)	402	50K v1	NAmr
United States (Wisconsin)	University of Wisconsin (Madison)	438	50K v1	NAmr

<sup>1</sup>HD = high density; 50K = 50,000 SNP; v1 = version 1; v2 = version 2; in the Netherlands, 700 and 447 animals were genotyped with the 50K v1 and v2 chips, respectively.

<sup>2</sup>N/A = not available; EuroAus = Australia, Denmark, Germany, Ireland, the Netherlands, New Zealand, and the United Kingdom; NAmr = United States, Canada, and the same research herd from the United Kingdom.



to impute the genotypes because it has been found to perform well compared with other methods when close relatives are not in the data set (Nicolazzi et al., 2013). The BEAGLE default of a total of 10 iterations was run for each chromosome.

A genomic relationship matrix (**GRM**) was calculated using the method of VanRaden (2008), with modifications as in Yang et al. (2010) for (1) the imputed genotypes of the UK animals after imputation from low to high density using 2 reference populations and (2) the genotypes of all animals that were retained after editing. Only SNP with a MAF >0.005 were included in the GRM.

Principal components analysis was performed on the GRM that included all animals, defining research farms as groups, to investigate whether animals of different countries were genetically divergent. A measure of population divergence (**F<sub>ST</sub>**) was also calculated using the genotypes of each research farm, using the formula of Hedrick (2005).

The diagonal of the GRM is sometimes considered to be the inbreeding of an individual +1 (Bolormaa et al., 2013). The diagonal of the GRM was compared with the average genomic homozygosity of an individual, which was calculated as the proportion of homozygous SNP for each individual in the data set.

RESULTS

Quality Control of 50K Genotypes

Editing of the 50K genotypes resulted in 3.8 and 4.2% of animals being removed due to low call rates for data sets EuroAus and NAmer, respectively, and a further 4.3 and 4.0% being removed due to pedigree conflicts (Table 2). The proportion of genotypes (by animal) retained for imputation to HD were similar for data sets EuroAus and NAmer (92 and 91%, respectively; Table 2).

In data set EuroAus, the number of useful 50K SNP retained by country varied between 37,732 and 44,670 (Table 3). In data set NAmer, the number of useful SNP retained by country varied between 38,225 and 51,141 (Table 4). Low call rate and missing heterozygotes when both homozygotes were present were the main reasons why SNP were deleted (Tables 3 and 4). In data set EuroAus, a similar proportion of SNP were removed through the editing process for both Illumina version 1 and Illumina version 2, whereas in data set NAmer, less than 10% of SNP on Illumina version 1 were removed (UK genotyped animals), but between 12 and 31% were removed for Illumina version 2 (Table 3). The large number of Mendelian inconsistencies for genotypes of the University of Wisconsin data probably

reflects the fact that these genotypes were shared immediately after they became available, and pedigree errors detected by genotyping had not yet been corrected.

The final 50K data set of EuroAus consisted of 47,391 SNP and 4,097 animals. Of these animals, approximately 1,000 were male and included to improve the imputation of missing 50K genotypes that were removed during quality control. The males were removed before imputation from 50K to HD. Data set NAmer had 1,579 animals and varying numbers of SNP per research herd were available for imputation to HD. The number of animals with sires that had HD genotypes (Table 5) varied between 0 (Denmark) and 325 (UK).

Quality Control of HD Genotypes

The initial HD data set for EuroAus consisted of 779,913 SNP for 3,313 animals. After removing animals with low call rates, 3,165 remained, of which 1,277 were male and the remainder were female from Australia and New Zealand (835 and 1,035, respectively). The males were registered for the first time by 11 different countries (Table 6). After removing SNP with low call rates (<95%) and GenTrain scores of less than 60%, 608,416 SNP remained. Other SNP were removed because they were on the Y chromosome, they were mismatched or unmapped or mitochondrial, or Mendelian inconsistencies were detected. These edits accounted for only 3% of the final SNP and the final number of SNP was 591,621. The genotypes that matched the SNP positions of data set EuroAus were selected from data set NAmer. The correlation between allele frequencies of HD genotypes in both data sets was 0.98.

Correlation of Allele Frequencies of Imputed Genotypes

The correlation between allele frequencies of the genotypes that were imputed from 50K to HD in data sets EuroAus and NAmer was very high (>0.98), although some SNP still differed considerably in allele

**Table 2.** Animals deleted from the genotype data following the quality control steps applied to data sets EuroAus (Australia, Denmark, Germany, Ireland, the Netherlands, New Zealand, and the United Kingdom) and NAmer (United States, Canada, and the same research herd from the United Kingdom)

Item	Data set [no. (%)]	
	EuroAus	NAmer
Received	4,460	1,735
Removed due to low call rate	170 (3.8)	73 (4.2)
Removed due to pedigree conflict	193 (4.3)	69 (4.0)
Remaining after editing	4,097 (92)	1,579 (91)

**Table 3.** The number of useful SNP following removal due to low call rates and Mendelian inconsistencies that were available for imputation to high density in data set EuroAus [Australia, Denmark (DNK), Germany, Ireland (IRL), the Netherlands, New Zealand, and the United Kingdom]

Item	the Netherlands		IRL		United Kingdom		DNK		Germany		
	Subset 1	Subset 2	Subset 1	Subset 2	Subset 1	Subset 2	Subset 1	Subset 2	Subset 1	Subset 2	Subset 3
SNP on the panel	54,001	54,609	54,001	54,001	54,001	54,001	54,609	54,001	54,001	54,001	54,609
Version of 50K panel <sup>1</sup>	1	2	1	1	1	1	2	1	1	1	2
Deleted: low call rate or homozygote/heterozygote <sup>2</sup>	9,104	12,647	11,175	16,042	10,119	16,042	10,199	11,753	13,303	13,303	14,880
Set to missing: Mendelian inconsistencies <sup>3</sup>	227	227	227	227	227	227	227	227	227	227	227
Useful SNP	44,670	41,735	42,599	37,732	43,655	37,732	44,183	42,021	40,471	39,502	39,502

<sup>1</sup>50K = 50,000 SNP.

<sup>2</sup>This includes also 620 SNP that were deleted because they had fewer than 500 genotypes left across all subsets after all quality control steps.

<sup>3</sup>This is the number of deleted SNP due to >2.5% Mendelian inconsistencies in data set EuroAus (including all subsets) across all SNP that survived all other quality control steps.

frequency when comparing the 2 data sets. This could either be because of genuine genetic differences between the 2 populations or because of errors arising from the imputation process. The correlation between allele frequencies calculated using imputed genotypes of 583 UK animals that were in both data sets was, not surprisingly, even stronger (>0.99).

On average, 7 and 21% of cows in data sets EuroAus and NAmr, respectively, were sired by bulls that were in the reference population of the respective populations (Table 5). A total of 1,277 bulls were in the reference population of data set EuroAus, of which around half (588) were also present in the HD reference populations of data set NAmr. So, the populations used to impute from 50K to HD were not independent and some similarity in imputation would arise as a consequence of having these bulls in common.

### Comparison of Imputation Accuracy Using 2 Reference Populations

No scenarios existed where known 700K genotypes were imputed after being set to missing. If there were, the realized correlation between the true and predicted genotype could have been calculated. However, 50K genotypes from one set of animals (from the UK) were imputed to HD using the reference population of both data sets EuroAus and NAmr. This allowed the proportion of imputed genotypes that differed as a result of the HD reference population used to be calculated. The average number of imputed SNP that were discordant (i.e., differed according to the reference population used) was 12,066 per animal (2.2% of the genotypes; Figure 1). The range was 41,571 (7%) discordant genotypes for the animal with the most discordant SNP down to 2,862 (0.4%) for the animal with the least discordant SNP, which could be due to their relationship to both reference populations.

The BEAGLE program does not explicitly use pedigree information in the imputation process and, in fact, the size of the reference population and genomic similarity to the reference population are likely to be more important. However, it is still possible that having close relatives, such as sires, in the reference population may be beneficial to imputation. Therefore, the average number of discordant SNP for cows that were sired by bulls that had HD genotypes in both data sets was compared. Two hundred and seventy cows were sired by 40 bulls that were genotyped at HD in both data sets. The sires of 313 cows were not genotyped at HD in either data set. The mean number of discordant SNP for animals with genotyped sires was slightly lower than animals with nongenotyped sires (12,399 and 12,505 respectively).

**Table 4.** The number of useful SNP following removal due to low call rates and Mendelian inconsistencies that were available for imputation to high density in data set NAmr [United States, Canada, and the same research herd from the United Kingdom (UK) as used in the European and Australasian data set]

Item	Canada	US (Wisconsin)	US (Iowa)	UK
Version of 50K panel <sup>1</sup>	2B	2B	2B	1
SNP on the panel	55,647	55,647	55,647	54,001
Deleted: low call rate or homozygote/heterozygote	8,834	9,081	6,723	2,452
Set to missing: Mendelian inconsistencies	1,096	8,887	80	408
Useful SNP	45,717	38,225	48,844	51,141

<sup>1</sup>50K = 50,000 SNP.

For one SNP, almost half (48%) of the genotypes of UK animals were imputed differently using EuroAus and NAmr HD reference data sets (Figure 2). In fact, 408 SNP had at least 20% of genotypes that were not in agreement between data sets (Figure 2). As these were considered to be poor-quality SNP, they were removed.

On average, the 2 data sets had 0.025% opposing homozygotes, so most of the discordant genotypes were from heterozygous to homozygous SNP and vice versa. The proportion of discordant SNP was strongly associated with the difference in allele frequencies between the 2 data sets of imputed genotypes of the UK animals (Figure 3). For the SNP with  $\geq 20\%$  discordant SNP, the difference between allele frequencies of the UK (imputed) genotypes was 0.07, whereas the difference between allele frequencies of data sets EuroAus and NAmr (HD genotypes) was 0.30. This result implies that large differences in allele frequencies of the reference population may have led to the discordant genotypes. However, these SNP also had poorer BEAGLE coefficients of determination. The SNP that were discordant for 20%+ of animals had a mean BEAGLE coefficient of determination of 79.7% [i.e., 17% lower than the mean coefficient of determination for all genotypes of UK animals, which was 96.5% (data set EuroAus only, data set NAmr animals were all imputed together)]. The mean difference between the allele dosage and the

most likely genotype of the SNP was calculated for each SNP. As expected, the differences were larger (10 $\times$ , on average) for the 446 SNP that had high ( $>0.2$ ) proportions of discordant genotypes.

Regions existed across the genome where differences between the genotypes that were imputed in data sets EuroAus and NAmr were more likely to arise (Figure 4); one example, on chromosome 10 spanned 520,948 bp. Here, 52 of the 103 SNP had  $\geq 20\%$  discordant SNP. Of the 52 SNP, the average proportion of animals that had discordant genotypes in this region was high (41%).

### Final Genomic Data Set

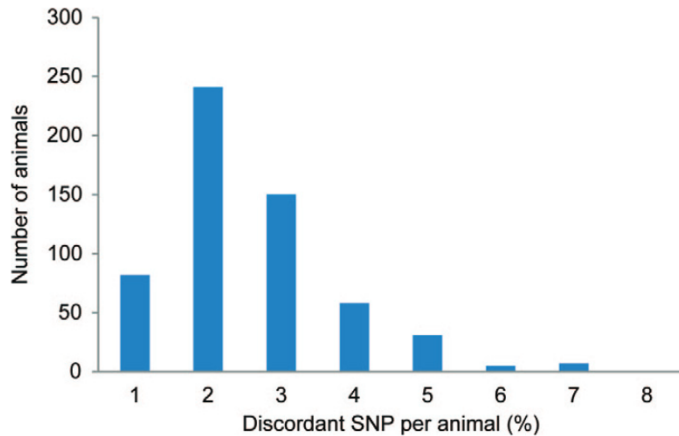
The final step was to combine the imputed genotypes from data sets EuroAus and NAmr. As the proportion of discordant SNP was slightly lower for animals with genotyped sires, imputed HD genotypes for UK animals from the EuroAus data set were selected (instead of from NAmr) in addition to the remaining genotypes from data sets EuroAus and NAmr. After editing, there were 591,213 genotypes for 5,999 animals from 10 research herds (the number of animals in each data set are in parentheses): Australia (835), Denmark (189), Canada (174), Germany (806), Ireland (399), the Netherlands (1,078), New Zealand (1,053), UK (655), United States (Iowa; 397), and United States (Wisconsin; 413).

**Table 5.** The number of animals whose genotypes were imputed from 50,000 (50K) SNP to high density (HD) and the number of animals with sires, maternal grandsires (MGS), and paternal grandsires (PGS) that were genotyped using the Illumina HD SNP chip (Illumina Inc., San Diego, CA)

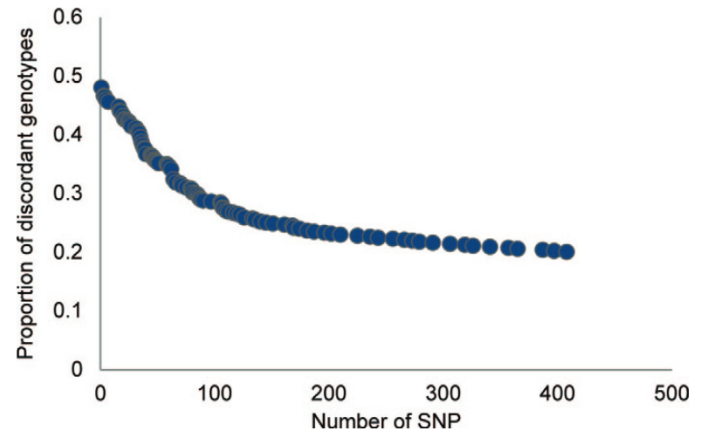
Country (state) <sup>1</sup>	Data set	No. of animals	No. with genotyped sire	No. with genotyped MGS	No. with genotyped PGS
CAN	NAmr	174	20	45	127
DEU	EuroAus	806	223	270	18
DNK	EuroAus	189	0	3	11
IRL	EuroAus	399	304	0	43
NLD	EuroAus	1,078	240	230	77
UK	NAmr	618	305	0	114
UK	EuroAus	600	325	300	40
US (IA)	NAmr	397	89	67	222
US (WI)	NAmr	413	4	1	4

<sup>1</sup>CAN = Canada; DEU = Germany; DNK = Denmark; IRL = Ireland; NLD = the Netherlands; UK = United Kingdom; IA = Iowa State University (Ames), WI = University of Wisconsin (Madison).





**Figure 1.** The percentage of discordant SNP (per animal) for 583 Holstein cows from the United Kingdom (UK) that were imputed from low [50,000 (50K)] to high [600,000 (600K)] SNP density in 2 separate data sets [EuroAus (Australia, Denmark, Germany, Ireland, the Netherlands, New Zealand, and the UK) and NAmers (United States, Canada, and the same research herd from the UK)]. Color version available in the online PDF.



**Figure 2.** The proportion of discordant SNP for 583 United Kingdom (UK) animals imputed in 2 separate data sets [EuroAus (Australia, Denmark, Germany, Ireland, the Netherlands, New Zealand, and the UK) and NAmers (United States, Canada, and the same research herd from the UK)]. The SNP with the most discordant genotypes had 48% of imputed genotypes that differed between data sets, and 408 SNP had more than 20% discordant genotypes. Color version available in the online PDF.

### Genomic Characterization of 10 Research Populations

In Figure 5, the first, second and third largest axes of variation are shown, obtained by principal components analysis of the GRM that included all animals. The first 3 eigenvalues accounted for 2.2, 1.0, and 0.8% of the variation. All 3 axes clearly differentiated the New Zealand heifers from the remaining population. When the New Zealand animals were excluded from the analysis, no further observable differentiation ex-

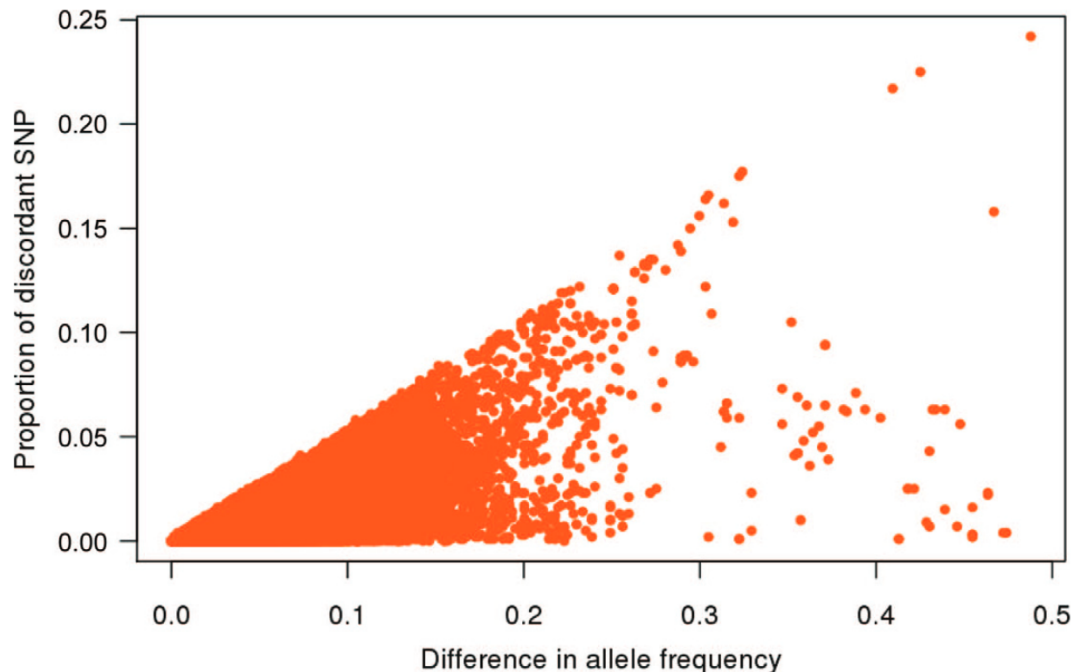
isted between the populations (results not shown); the sum of the first 3 eigenvalues of the GRM excluding New Zealand was 2.9%, which is around 1% less than the GRM that included all genotypes. Thus, animals from the remaining populations were genetically more similar.

The mean  $F_{ST}$  (multiplied by 100) of genotypes from each country, when compared with all the other genotypes combined, varied between 0.23 (Australia) and 1.47 (New Zealand; Table 7). So, on average, the New

**Table 6.** Number of bulls by country of origin of bulls in the reference populations of data sets EuroAus [Australia, Denmark, Germany, Ireland, the Netherlands, New Zealand, and the United Kingdom (UK)] and NAmers (United States, Canada, and the same research herd from the UK) used to impute 50,000 (50K) SNP genotypes to high density (HD) in both data sets and the number of sires, maternal (MGS), and paternal (PGS) grandsires of United Kingdom (UK) animals that were in both data sets

Country of origin <sup>1</sup>	Data set		Number of bulls in common	Number of bulls in common (UK)		
	EuroAus	NAmers		Sire	MGS	PGS
AUS	79	2				
CAN	114	110	81	9	15	3
CHE		1				
DEU	53	63	23			
DNK	9	7	3			
ESP		2				
FRA	85	36	16	1	1	
IRL	44	7				
ITA	28	709	20	1	1	
NLD	272	140	61	9	9	1
NZL	38	6	5			
UK	337	400	245	11	13	
US	218	535	134	9	10	9
Total	1,277	2,018	588	40	49	13

<sup>1</sup>AUS = Australia; CAN = Canada; CHE = Switzerland; DEU = Germany; DNK = Denmark; ESP = Spain; FRA = France; IRL = Ireland; ITA = Italy; NLD = the Netherlands; NZL = New Zealand.



**Figure 3.** The proportion of discordant SNP for genotypes of 583 United Kingdom (UK) animals imputed in 2 separate data sets [EuroAus (Australia, Denmark, Germany, Ireland, the Netherlands, New Zealand, and the UK) and NAmr (United States, Canada, and the same research herd from the UK)] plotted against the difference in imputed allele frequencies (the correlation between allele frequency and proportion of discordant SNP was 0.75). Color version available in the online PDF.

Zealand population was the most divergent, confirming the results of the principal components analysis.

The mean proportion of homozygous SNP of individuals was 0.64 and was similar for each country (ranging from 0.640 in New Zealand to 0.653 in Canada). However, for the same value of the diagonal element of the GRM, the proportion of SNP that were homozygous was lower for the New Zealand and the Irish populations (which also had a low mean proportion of homozygous SNP) compared with other populations (Figure 6).

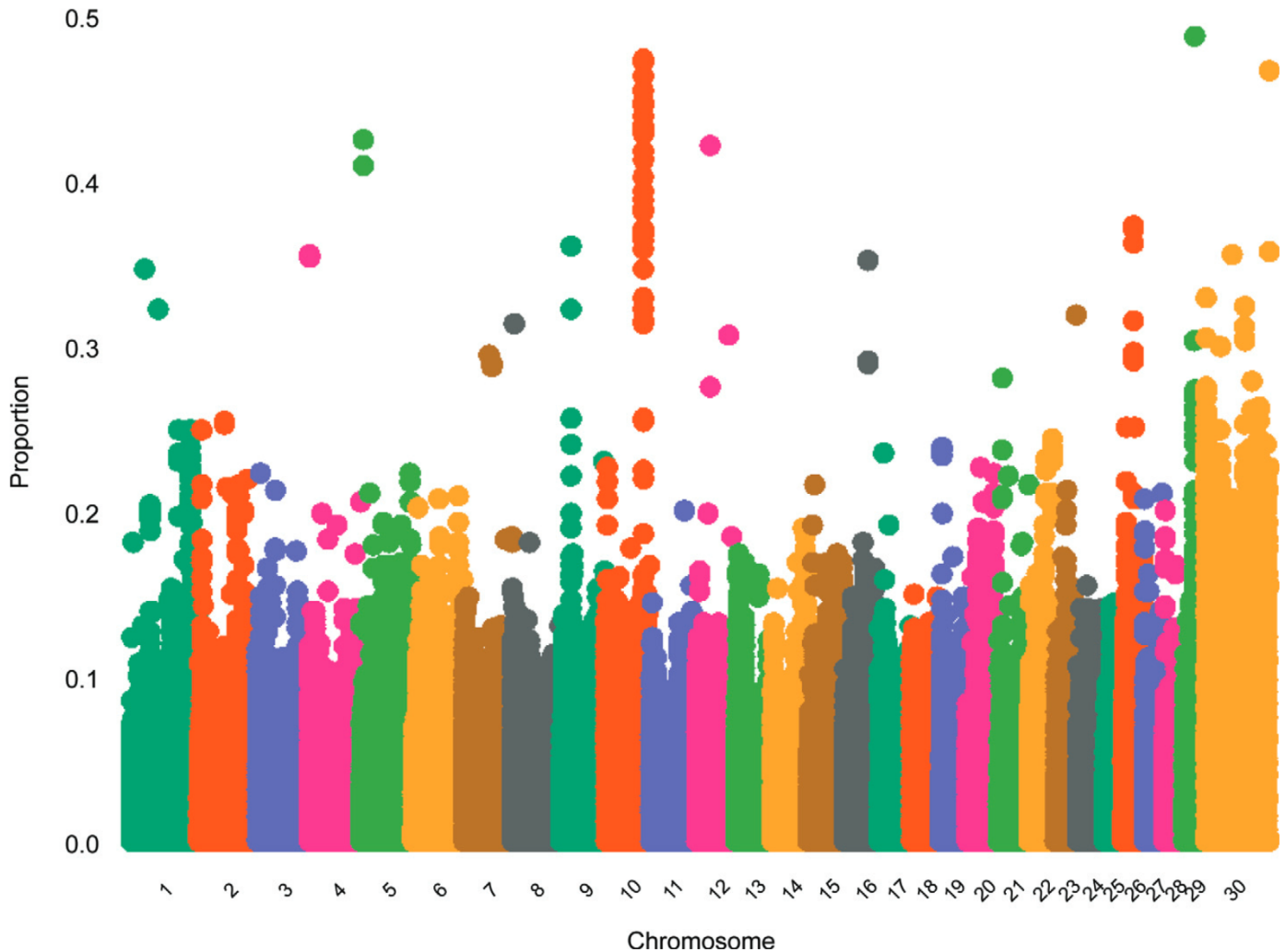
## DISCUSSION

High-density genotypes have been documented to increase the accuracy of genomic prediction using data from research herds (Pryce et al., 2012) and have also been reported to increase the precision of QTL detection in GWAS (VanRaden et al., 2013). For these reasons, imputing genotypes to high density in research herds is likely to be desirable, especially as these data sets often have fewer animals with valuable phenotypes that are not generally measured in national dairy populations (e.g., DMI). In some situations, it is not practical to impute the genotypes of animals in 1 data set, because of restrictions imposed on sharing of genotypes; although it is unusual to have 2 reference populations for imputation,

as was the case in the present study, it can be used to an advantage, as SNP where a large proportion of animals are imputed differently can be identified and removed.

In combining genotypes from several research herds, ideally the same quality control process to genotype editing should be applied to each data set to be combined. Although the genotypes used in this study originated from 10 research populations, the genotype editing and imputation steps were performed in 2 independent data sets. As was expected, similar proportions of animals (around 4%) were removed from each of the 2 data sets because of low call rate and because of pedigree conflicts (Table 2). To compare the imputation of data sets EuroAus and NAmr, 1 group of genotypes (from UK animals) was imputed as part of both data sets. The proportion of discordant SNP in the UK population that was included in both data sets that underwent imputation was 2.2%.

The majority of discordant genotypes were switches from heterozygotes to homozygotes and vice versa; very few (0.025%) were imputed as one homozygote in one data set and the opposing homozygote in the other imputation. Furthermore, it appeared that the proportion of discordant SNP was strongly associated with difference in allele frequency between data sets (Figure 3). So, if populations are imputed separately



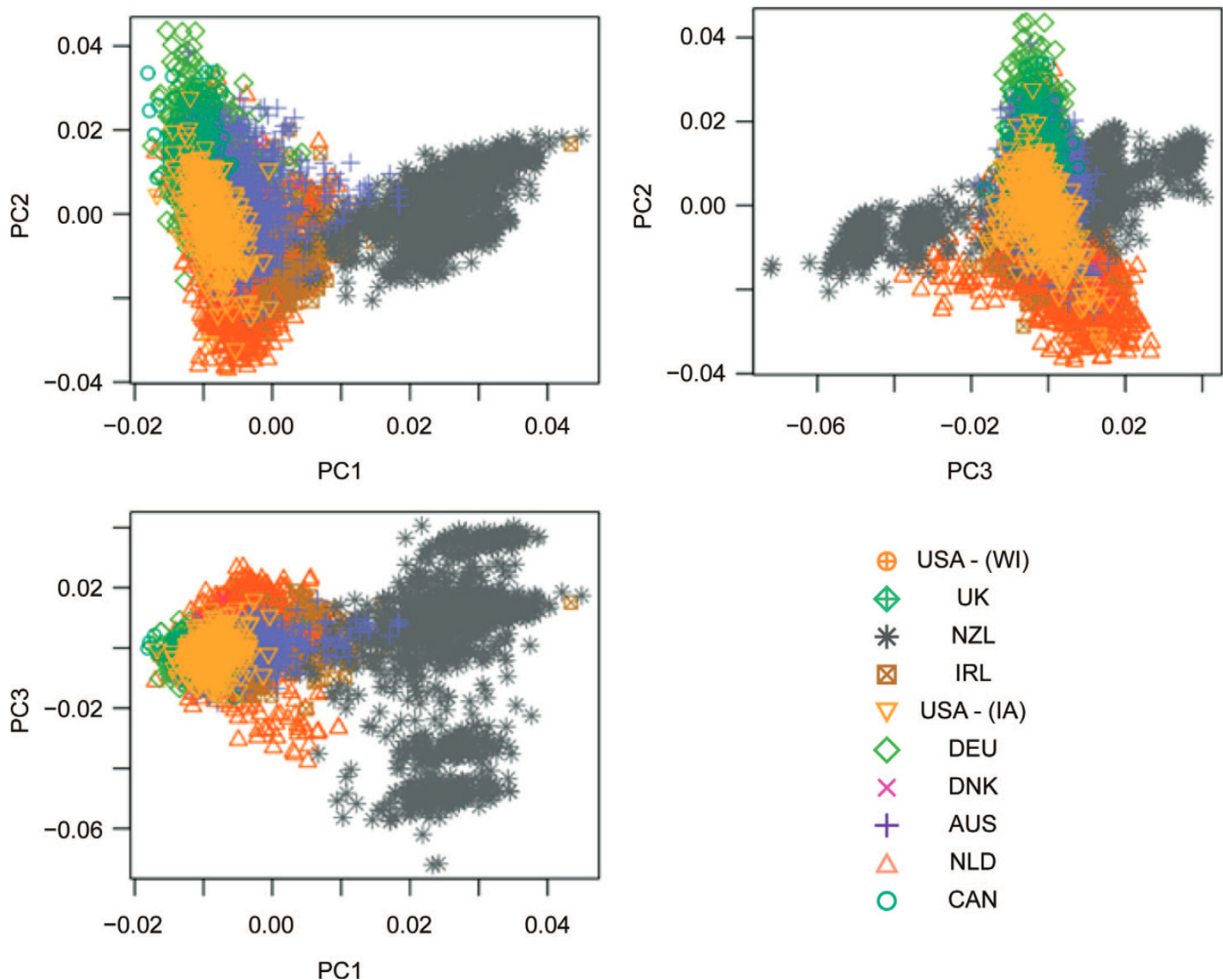
**Figure 4.** The proportion of United Kingdom (UK) animals ( $n = 583$ ) with genotypes that were imputed differently in data sets EuroAus (Australia, Denmark, Germany, Ireland, the Netherlands, New Zealand, and the UK) and NAmr (United States, Canada, and the same research herd from the UK) across the genome. Color version available in the online PDF.

and later combined, it may be advisable to exclude SNP that have large differences in allele frequency. The limitation of this approach is that some SNP will be excluded unnecessarily (in Figure 3, some SNP have a low proportion of discordant SNP and large differences in allele frequency). Alternatively, the mean difference between the allele dosage and the most likely genotype can be used to exclude SNP, which was 10 times larger for the 446 most discordant SNP than other imputed SNP.

It is important to note that not all discordant SNP are errors; the imputation is likely to be more accurate for an individual that is more closely related to reference animals in one data set (compared with the other data set). On average, 7 and 21% of cows in data sets EuroAus and NAmr, respectively, were sired by bulls that were in the reference population of the respec-

tive populations (Table 5). Furthermore, out of 583 UK cows that had genotypes imputed as part of both data sets (EuroAus and NAmr) 270 cows were sired by bulls that had HD genotypes in both data sets. The mean number of discordant SNP for animals with genotyped sires was slightly lower than animals with non-genotyped sires: out of 591,621 SNP, 12,399 and 12,505 were discordant SNP for animals with genotyped and nongenotyped sires, respectively. However, the difference is not large enough to conclude that having a sire in the data set improved imputation accuracy. However, Mulder et al. (2012) and Zhang and Druet (2010) found that the accuracy of imputation was influenced by the relationship between animals genotyped at low density [ $<10,000$  (10K) SNP] and high density (50K); animals that had both parents genotyped at 50K had fewer imputation errors. It is possible that having a





**Figure 5.** The first 3 principal components (PC) of the genomic relationship matrix constructed using high-density genotypes from 9 countries and 10 research populations [United States (USA): University of Wisconsin (WI), United Kingdom (UK), New Zealand (NZL), Ireland (IRL), USA: Iowa State University (IA), Germany (DEU), Denmark (DNK), Australia (AUS), the Netherlands (NLD), and Canada (CAN)]. Together, the first 3 principal components accounted for 3.9% of the variance. Color version available in the online PDF.

close (genetic/genomic) relationship to the reference population is more important for imputation from low density to 50K than from 50K to 700K. The accuracy of imputation increases as the genetic diversity within the reference population increases (Hayes et al., 2012) because more of the haplotypes in the genotypes under imputation are likely to exist. The accuracy of imputation within a population appears to increase further still by using an even larger and diverse population genotyped at high density. For example, Brøndum et al. (2012) found that the accuracy of imputation of Danish, Swedish, or Finnish Red populations could be increased by including Holsteins in the reference population, implying that a larger reference population

that captures more of the LD between markers may be beneficial in improving imputation.

The combined imputed genotype data sets described in the current study have already been used to estimate genetic correlations of DMI measured in several countries (Berry et al., 2013) and they will also be available as a resource in future studies (for example, estimating the accuracy of genomic prediction of DMI using a multi-country reference population). They could also be used for GWAS where the aim is to identify regions across the genome that affect DMI. To minimize the effect of poorly imputed regions on future GWAS, SNP that were in regions of high (>20%) discordant SNP were removed as part of the quality control process.

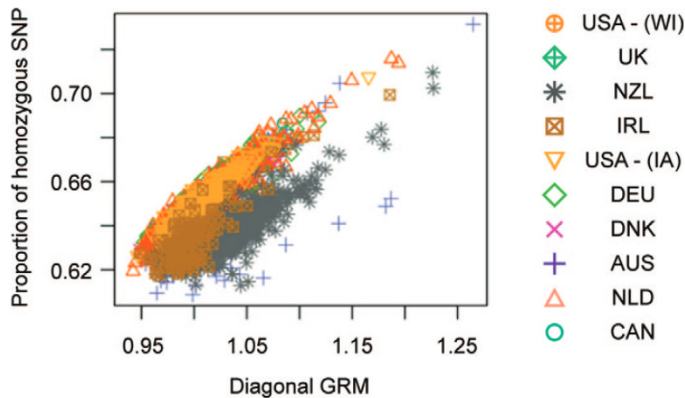


Figure 6. Proportion of homozygous SNP per animal compared with the diagonal elements of the genomic relationship matrix (GRM) constructed using high-density genotypes from 9 countries and 10 research populations [United States (USA): University of Wisconsin (WI), United Kingdom (UK), New Zealand (NZL), Ireland (IRL), USA: Iowa State University (IA), Germany (DEU), Denmark (DNK), Australia (AUS), the Netherlands (NLD), and Canada (CAN)]. Color version available in the online PDF.

The threshold of percentage of discordant SNP was arbitrarily defined. However, as Figure 4 shows, using this threshold will capture some regions where mapping (or other) errors may have occurred, as discordant SNP were likely to be close together on the genome.

One region on chromosome 10 had more than 20% discordant SNP for 52 out of 103 SNP. The mean BEAGLE coefficient of determination in this region was 45%, indicating that this region was poorly imputed. One possible cause could have been weak LD between markers, which would indicate that the construction of the SNP map in this region was poor. However, this did not appear to be the case, as the average pairwise LD measured using the squared correlation coefficient ( $r^2$ ; Hill and Robertson, 1968) was 0.16 for the 103-SNP region with a high number of discordance, whereas it was 0.06 for 100-SNP chunks across the whole of chromosome 10. Furthermore, sire-son concordance for SNP in this region has been evaluated using sequence data (H. Daetwyler, Department of Environment and Primary Industries, Bundoora, VIC, Australia, personal communication, 2013) and again, there is no indication of poor mapping in this region.

One of the limitations of the study was that the 700K map used for imputation of genotypes from 50K to 700K for both data sets (EuroAus and Namer) was constructed by applying quality control processes to the 700K SNP of the EuroAus data set only. This could have led to some SNP that were of poor quality in data set Namer and may be the reason for poorer imputation in some regions. This was one of the reasons why the imputation results of data set EuroAus were

Table 7. Mean, SD, and maximum measure of population divergence ( $F_{ST}$ ), calculated using genotypes for each country when compared with all the other countries combined

Country <sup>1</sup>	Mean × 100	SD × 100	Maximum × 100
AUS	0.23	0.35	6.2
CAN	0.82	1.13	17.6
DEU	0.43	0.60	23.2
DNK	0.60	0.82	9.5
IRL	0.39	0.55	8.2
NLD	0.33	0.45	7.0
NZL	1.47	2.00	28.9
UK	0.27	0.40	6.4
US (IA)	0.40	0.57	7.9
US (WI)	0.41	0.57	14.9

<sup>1</sup>AUS = Australia; CAN = Canada; DEU = Germany; DNK = Denmark; IRL = Ireland; NLD = the Netherlands; NZL = New Zealand; UK = United Kingdom; IA = Iowa State University (Ames); WI = University of Wisconsin (Madison).

selected for UK animals. The other reason was that the population genotyped at HD was more closely related to the UK 50K population in data set EuroAus (Table 5).

The GRM was also used to assess the genetic diversity among the research populations. This is useful in assessing the genetic similarity between populations and may be indicative of the likely success of genomic prediction in the different populations (Habier et al., 2007; Pszczola et al., 2012). The principal components analysis (Figure 5) and  $F_{ST}$  results showed that the New Zealand population was consistently genetically different from the other populations of cattle, in accordance with de Roos et al. (2008) and Pryce et al. (2012). Furthermore, more genetic diversity existed within the New Zealand population, which could partly be a result of the New Zealand population being graded up from Jerseys and because the population is an admixture of cattle of New Zealand and North American ancestry (Harris and Kolver, 2001). This could also be why the mean proportion of homozygous SNP was slightly lower for New Zealand compared with other populations (Figure 6). A low proportion of homozygous SNP is likely to be indicative of lower inbreeding. However, the diagonal of the GRM, sometimes used as a measure of inbreeding, was actually, on average, higher for the New Zealand population. The average allele frequencies used to calculate the GRM will be dominated by larger populations in the data set. This would lead to inflation of the diagonal elements of the GRM for New Zealand. Therefore, unless the GRM is constructed by correcting for differences in average allele frequencies that arise in multiple populations (Erbe et al., 2012), the diagonal elements of the GRM are only indicative of inbreeding within population.



For other research populations considering combining genotypes, the following guidelines should be considered:

- Ideally, a single reference population of HD genotypes should be used for imputation, as this is expected to lead to the highest accuracy of imputation.
- If it is not possible to impute genotypes together,
- Include a subset of genotypes in all imputation data sets, so that the proportion of discordant SNP in each imputation set can be compared;
- Consistent genotyping editing rules need to be applied across subsets of data;
- Exclude SNP where a large proportion of discordant SNP arise, or alternatively exclude SNP where the difference in allele frequencies is large between data sets.

## CONCLUSIONS

Establishing a genomic reference population for difficult-to-measure traits such as DMI can be achieved by combining data sets from research herds including genotyped cows that have records for such traits. This procedure may involve imputation of the genotypes to high-density SNP to make optimal use of the data. Results from this study demonstrated that if genotypes of several research herds are to be imputed to HD separately and later combined, it is useful to include a set of low-density genotypes in each of the imputation data sets, so that the proportion of discordant genotypes can be computed for each imputed SNP. Discordant genotypes can be used to identify poor-quality SNP that can be removed from the combined data set. If imputation has to be done separately with no overlap in genotypes, then large differences in allele frequencies between imputed genotypes may be used to identify and remove poorly imputed SNP. Finally, a GRM constructed using the genotypes from the 10 populations showed greater genomic diversity within the New Zealand population itself and between New Zealand and other populations. Accordingly, average SNP homozygosity was also lower in the New Zealand population.

## ACKNOWLEDGMENTS

This work was performed by the global Dry Matter Initiative (gDMI) consortium in which 15 parties from 11 countries combined their knowledge, data, and expertise. Financial support from CRV (Arnhem, the Netherlands), Irish Cattle Breeding Federation (ICBF; Bandon, Co. Cork, Ireland), Confederacion de Asociaciones de Frisona Española (CONAFE; Madrid,

Spain), DairyCo (Warwickshire, UK), and DairyGen Council of Canadian Dairy Network (Guelph, ON, Canada) is gratefully appreciated, as well as the EU FP7 International Research Staff Exchange Scheme (IRSES) SEQSEL (grant no. 317697). The providers of genotypes used in this study are gratefully acknowledged: Department of Environment and Primary Industries (DEPI; Melbourne, VIC, Australia), Dairy Futures CRC (Melbourne, VIC, Australia), University of Alberta (Edmonton, AB, Canada), Aarhus University (Tjele, Denmark), Christian-Albrechts-University (Kiel, Germany), LIC and DairyNZ (Hamilton, New Zealand), the RobustMilk project (Lelystad, the Netherlands; Irish, part of the UK, and part of Dutch genotypes), the National Institute of Food and Agriculture [NIFA, Washington, DC; Iowa State University (Ames), University of Wisconsin (Madison), and part of the Netherlands and UK genotypes], and the Scottish Government (part of the UK genotypes). The Scotland's Rural College (SRUC; Edinburgh, UK) HD genotypes were made available as part of the UK Ruminant Genetic Improvement Network, funded by the Department for Environment, Food and Rural Affairs (Defra, London, UK; IF0169). The genotyping of Danish animals was done as part of the project "Genomic Selection—From function to efficient utilization in cattle breeding," (grant no. 3405-10-0137). The genotyping of German animals was financially supported by the German Federal Ministry of Education and Research (Berlin, Germany; project FUGATO-plus GENOTRACK, grant no. 0315134A), the Kompetenzzentrum Milch – Schleswig-Holstein (KMSH; Kiel, Germany), and the Nord-Ost Genetic GmbH & Co. KG (NOG; Verden, Germany). The Australian research was funded by the Gardiner Foundation (Melbourne, VIC, Australia) and Dairy Futures CRC. The NIFA grant no. 0224899 was funded by the US Department of Agriculture (USDA). This research was partially supported by Agriculture and Food Research Initiative Competitive Grants no. 2008-35205-18711 and 2011-68004-30340 from the USDA National Institute of Food and Agriculture. The RobustMilk project was financially supported by the European Commission under the Seventh Research Framework Programme, Grant Agreement KBBE-211708. Finally, thanks to Bolormaa Sunduimijid (Department of Environment and Primary Industries, Melbourne, Australia) for the use of her program to calculate  $F_{ST}$ .

## REFERENCES

- Banos, G., M. P. Coffey, R. F. Veerkamp, D. P. Berry, and E. Wall. 2012. Merging and characterising phenotypic data on conventional and rare traits from dairy cattle experimental resources in three countries. *Animal* 6:1040–1048.



- Berry, D. P., J. W. M. Bastiaansen, R. F. Veerkamp, S. Wijga, E. Wall, B. Berglund, and M. P. L. Calus. 2012. Genome-wide associations for fertility traits in Holstein-Friesian dairy cows using data from experimental research herds in four European countries. *Animal* 6:1206–1215.
- Berry, D. P., M. P. Coffey, J. E. Pryce, Y. de Haas, P. Lovendahl, N. Krattenmacher, J. J. Crowley, D. Spurlock, K. A. Weigel, K. A. Macdonald, and R. F. Veerkamp. 2013. International genetic evaluations for feed intake in dairy cattle. *Interbull Bull.* 47. Interbull, Uppsala, Sweden.
- Bolormaa, S., J. E. Pryce, K. Kemper, K. Savin, B. J. Hayes, W. Barendse, Y. Zhang, C. M. Reich, B. A. Mason, R. J. Bunch, B. E. Harrison, A. Reverter, R. M. Herd, B. Tier, H.-U. Graser, and M. E. Goddard. 2013. Accuracy of prediction of genomic breeding values for residual feed intake, carcass and meat quality traits in *Bos taurus*, *Bos indicus*, and composite beef cattle. *J. Anim. Sci.* 91:3088–3104.
- Brøndum, R. F., P. Ma, M. S. Lund, and G. Su. 2012. Short communication: Genotype imputation within and across Nordic cattle breeds. *J. Dairy Sci.* 95:6795–6800.
- Browning, B. L., and S. R. Browning. 2009. A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84:210–223.
- Cole, J. B., P. M. VanRaden, J. R. O'Connell, C. P. Van Tassell, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and G. R. Wiggans. 2009. Distribution and location of genetic effects for dairy traits. *J. Dairy Sci.* 92:2931–2946.
- de Haas, Y., M. P. L. Calus, R. F. Veerkamp, E. Wall, M. P. Coffey, H. D. Daetwyler, B. J. Hayes, and J. E. Pryce. 2012. Improved accuracy of genomic prediction for dry matter intake of dairy cattle from combined European and Australian data sets. *J. Dairy Sci.* 95:6103–6112.
- de Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of genomic predictions across multiple populations. *Genetics* 183:1545–1553.
- de Roos, A. P. W., B. J. Hayes, R. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein Friesian, Jersey and Angus cattle. *Genetics* 179:1503–1512.
- Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* 95:4114–4129.
- Gao, H., G. Su, L. Janss, Y. Zhang, and M. S. Lund. 2013. Model comparison on genomic predictions using high-density markers for different groups of bulls in the Nordic Holstein population. *J. Dairy Sci.* 96:4678–4687.
- Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177:2389–2397.
- Harris, B. L., F. E. Creagh, A. M. Winkelman, and D. L. Johnson. 2011. Experiences with the Illumina high density bovine Bead-Chip. *Interbull Bull.* 44:3–7.
- Harris, B. L., and E. S. Kolver. 2001. Review of holsteinization on intensive pastoral dairy farming in New Zealand. *J. Dairy Sci.* 84(E. Suppl.):E56–E61.
- Hayes, B. J., P. J. Bowman, H. D. Daetwyler, J. W. Kijas, and J. H. J. van der Werf. 2012. Accuracy of genotype imputation in sheep breeds. *Anim. Genet.* 43:72–80.
- Hedrick, P. W. 2005. *Genetics of Populations*. 3rd ed. Page 489. Jones and Barlett Publishers Inc., Sudbury, MA.
- Hickey, J. M., B. P. Kinghorn, B. Tier, J. H. J. van der Werf, and M. A. Cleveland. 2012. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. *Genet. Sel. Evol.* 44:9.
- Hill, W. G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38:226–231.
- Ma, P., R. F. Brøndum, Q. Zhang, M. S. Lund, and G. Su. 2013. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. *J. Dairy Sci.* 96:4666–4677.
- Marchini, J., and B. Howie. 2010. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11:499–511.
- Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS ONE* 4:e5350.
- Mulder, H. A., M. P. L. Calus, T. Druet, and C. Schrooten. 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. *J. Dairy Sci.* 95:876–889.
- Nicolazzi, E. L., S. Biffani, and G. Jansen. 2013. Short communication: Imputing genotypes using PedImpute fast algorithm combining pedigree and population information. *J. Dairy Sci.* 96:2649–2653.
- Pryce, J. E., J. Arias, P. J. Bowman, S. R. Davis, K. A. Macdonald, G. C. Waghorn, W. J. Wales, Y. J. Williams, R. J. Spelman, and B. J. Hayes. 2012. Accuracy of genomic predictions of residual feed intake and 250-day bodyweight in growing heifers using 625,000 single nucleotide polymorphism markers. *J. Dairy Sci.* 95:2108–2119.
- Pszczola, M., T. Strabel, H. A. Mulder, and M. P. L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. *J. Dairy Sci.* 95:389–400.
- Sargolzaei, M., J. Chesnais, and F. Schenkel. 2011. FImpute—An efficient imputation algorithm for dairy cattle populations. *J. Anim. Sci.* 89:(E-Suppl. 1):421. (Abstr.)
- Su, G., R. F. Brøndum, P. Ma, B. Guldbrandtsen, G. P. Aamand, and M. S. Lund. 2012. Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. *J. Dairy Sci.* 95:4657–4665.
- VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414–4423.
- VanRaden, P. M., D. J. Null, M. Sargolzaei, G. R. Wiggans, M. E. Tooker, J. B. Cole, T. S. Sonstegard, E. E. Connor, M. Winters, J. B. C. H. M. van Kaam, A. Valentini, B. J. Van Doormaal, M. A. Faust, and G. A. Doak. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. *J. Dairy Sci.* 96:668–678.
- VanRaden, P. M., J. R. O'Connell, G. R. Wiggans, and K. A. Weigel. 2011. Genomic evaluations with many more genotypes. *Genet. Sel. Evol.* 43:10.
- Veerkamp, R. F., M. P. Coffey, D. P. Berry, Y. de Haas, E. Strandberg, H. Bovenhuis, M. P. L. Calus, and E. Wall. 2012. Genome-wide associations for feed utilisation complex in primiparous Holstein-Friesian dairy cows from experimental research herds in four European countries. *Animal* 6:1738–1749.
- Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42:565–569.
- Zhang, Z., and T. Druet. 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. *J. Dairy Sci.* 93:5487–5494.